

Actas de las 4ª Jornadas de Intercambio y Reflexión acerca de la Investigación en
Bibliotecología, La Plata, 29-30 de octubre de 2015. La Plata: Facultad de
Humanidades y Ciencias de la Educación, Universidad Nacional de La Plata, 2015.
ISSN 1853-5631

Aplicaciones métricas de Google Scholar para la evaluación del impacto científico

Enrique Orduna-Malea^{1*}, Juan Manuel Ayllón², Alberto Martín-Martín² y Emilio Delgado López-Cózar²

¹ Universidad Politécnica de Valencia (España). Email: enorma@upv.es

² Universidad de Granada (España).

Resumen:

La aparición de Google Scholar en 2004 supuso una revolución tanto en los procesos de búsqueda de información académica como en los de evaluación científica (especialmente en las áreas de las Ciencias sociales y humanas) gracias a su ingente base de datos, basada en la recopilación de cualquier documento académico online, a pesar de los errores e imprecisiones existentes en la correcta vinculación de citas. Este trabajo presenta la elaboración de diferentes productos métricos de información elaborados a partir de Google Scholar (H Index Scholar, Publishers Scholar Metrics, Journal Scholar Metrics, La Biblioteconomía Española según Google Scholar Citations) con el propósito de mostrar la potencialidad y utilidad de esta base de datos a la hora de ser utilizada por los diferentes actores involucrados en la creación, difusión y evaluación de la actividad científica.

Palabras clave:

Motores de búsqueda académicos; Google Scholar, Evaluación científica; Bibliometría.

1. Introducción

Google Scholar es un motor de búsqueda académico oficialmente lanzado en noviembre de 2004 por la compañía de Mountain View, con apenas unas semanas de diferencia respecto al lanzamiento de Scopus, por parte de la multinacional Elsevier.

Scopus surgía con la clara intención de romper el monopolio que representaba el binomio Web of Science/Journal Citation Reports en el mercado de las bases de datos bibliográficas a partir de una herramienta conceptualmente muy similar (si bien con una interfaz gráfica diferente, una mayor cobertura de revistas y algunos indicadores bibliométricos diferentes). Google Scholar llegaba sin embargo representando unos objetivos manifiestamente diferentes; un mercado complementario: el de los motores de búsqueda académicos (Ortega, 2014).

En ese sentido, Google Scholar nacía como un buscador especializado en la indización y recuperación de material académico, de forma similar a como funcionaban otros motores

existentes como los pioneros Citeseer o Scirus, o las posteriores versiones de Microsoft Academic Search (Jacsó, 2011; Orduña-Malea et al, 2014).

Sin embargo, una de las prestaciones que incluía Google Scholar (y otros motores de búsqueda académicos) era la incorporación de datos de citas recibidas por los artículos indizados en la base de datos. Esta información, que suponía un claro criterio de filtrado de calidad para los usuarios, convertía a los motores de búsqueda académicos en buscadores con unas características peculiares, pues:

- Las citas son calculadas por el sistema de forma completamente automática. El programa identifica las referencias bibliográficas contenidas en cada artículo y las vincula o atribuye a los correspondientes artículos citados.
- Las citas provienen de la totalidad de los documentos indizados en el sistema, cuya cobertura sólo tiene limitaciones técnicas (documentos con resumen depositados en lugares de confianza, como repositorios, universidades, editoriales, catálogos bibliográficos, etc.). Por tanto, se basan en la computación total de impacto en citas, y no solo de la proveniente desde un grupo elitista de revistas seleccionadas a priori.

Estas características, unidas a la rápida expansión y dinamismo del sistema (actualizado diariamente), convirtieron pronto a Google Scholar en un producto con una cobertura amplísima. Es más, conforme crecía el tamaño, la precisión de los resultados de búsqueda aumentaba igualmente. En efecto, la cobertura de Google Scholar cubre todas las tipologías documentales (desde artículos de revistas de impacto hasta tesis doctorales, libros, informes, presentaciones en congresos, patentes o material didáctico), lo que proporciona cierta luz a las disciplinas humanísticas y sociales, con unas dinámicas de publicación diferentes a las ciencias (fundamentadas en la publicación de artículos escritos en inglés en revistas).

Este hecho provocó que tanto los investigadores como las instituciones y organismos de evaluación centraran su interés en la evaluación de Google Scholar como herramienta válida en tareas relacionadas con la evaluación científica (Torres-Salinas et al, 2009).

La automatización de las citas bibliográficas provocaba errores inadmisibles en procesos evaluativos. Cabe destacar la obra de Peter Jacsó al respecto, quien llevó a cabo una serie de estudios empíricos en los que demostró importantes errores de Google Scholar en la descripción bibliográfica de los documentos indizados, que provocaban a su vez importantes errores en la atribución de autoría y de citas recibidas (Jacsó, 2008; 2010). Por el contrario, surgió otra corriente que trató de justificar la necesidad de Google Scholar como herramienta de evaluación en Ciencias humanas y sociales, dadas las insuficiencias y claros sesgos de las bases de datos bibliográficas hegemónicas (WoS y Scopus). Esta corriente, representada entre

otros por los trabajos de Anne-Wil Harzing (Harzing y Van der Wal, 2007), era consciente de los errores y limitaciones de Google Scholar, por lo que trataba de centrar su uso en la localización de citas únicas y no tanto en los datos totales de citas ofrecidos (Kousha y Thelwall, 2007).

2. La mirada bicéfala de Google Scholar

Durante los últimos 10 años, se ha producido un debate en la arena académica acerca del uso de Google Scholar en tanto que herramienta de evaluación científica. Y en ese sentido, para comprender adecuadamente este debate, conviene diferenciar las dos grandes caras que presenta el producto: Google Scholar como buscador versus Google Scholar como herramienta de evaluación bibliográfica.

El estudio de Google Scholar en tanto que buscador ha sido tradicionalmente abordado por el sector bibliotecario, desde donde se ha estudiado sus características, sus prestaciones de búsqueda, cobertura y precisión de resultados. Actualmente, pese a la escasez de funcionalidades avanzadas de búsqueda, existe cierto consenso hacia la eficacia en sus resultados, su rapidez, sencillez, facilidad de funcionamiento, amplia cobertura y, muy especialmente, su naturaleza universal y gratuita. Esto ha permitido que en la actualidad Google Scholar se haya convertido en el lugar favorito para comenzar los procesos de búsqueda de material académico tanto para estudiantes como para investigadores, quienes manifiestan un uso intensivo del buscador (Van Noorden, 2014).

El análisis de Google Scholar en tanto que herramienta de evaluación científica está siendo abordado desde el campo de la webmetría, bibliometría y cienciometría, donde la cobertura, evolución y precisión tanto de la información bibliométrica (Orduña-Malea et al, 2015) como de los indicadores obtenidos, tanto a nivel de autores como de revistas (Bar-Ilan, 2008; Harzing y Van der Wal, 2009; Jacsó 2009; Orduña-Malea y Delgado López-Cózar, 2014), están siendo actualmente objeto de análisis científico. Google Scholar ha entrado definitivamente en la agenda política; aquel buscador que en 2010 era rechazado de forma rotunda por parte del sector más ortodoxo de la disciplina ha pasado a ser considerado, no sin cierta reserva, y más por demanda popular que por interés. Por utilizar la conocida metáfora de Umberto Eco, la posición acerca del uso de Google Scholar pasa desde los apocalípticos a los integrados.

Sin entrar en debates ni opiniones, el hecho es que Google Scholar en 2015 es un producto muy diferente al de 2004. Por tanto, ante la pregunta de si Google Scholar (y su familia de productos) podría ser considerada como una fuente alternativa (o complementaria) para la evaluación científica, la respuesta en 2004 (y en 2010) eran de un rotundo no. Los autores manifestamos, a fecha de 2015, un sí condicional y empírico (no opinático).

En la actualidad Google Scholar presenta la base de datos bibliográfica más grande del mundo; en 2014 su tamaño se estimaba en 170 millones de documentos (Orduña-Malea et al, 2015). Además, su tasa de crecimiento es más grande que Web of Science o Scopus, aunque lo hace de manera irregular (Orduña-Malea y Delgado López-Cózar, 2013). Google Scholar constituye el primer producto de big data académico de la historia, abanderando la transición desde un sistema cerrado, controlado y basado en la contabilización de citas (entre revistas de elite) hacia otro abierto, no moderado y fundamentado en la contabilización de menciones desde y hacia cualquier recurso académico. La clave de esta aproximación es entender que la elite, el material de calidad, también se encuentra incluido dentro de Google Scholar, y que muchos materiales de relevancia se encuentran en Google Scholar, y no en WoS/Scopus.

El principal problema radica en que el sistema impide, dadas sus pocas prestaciones documentales, el correcto filtrado y exportación de datos para su análisis posterior. Por otra parte, la exclusiva dependencia de la Web limita su estabilidad (los materiales que desaparecen de la Web, desaparecen de la base de datos). Pese a estas grandes desventajas, el sistema ofrece la posibilidad de explorar nuevas dimensiones del concepto de impacto, tradicionalmente basado en las citas provenientes de ciertas revistas, pero obviando el resto (es decir, fiándose exclusivamente de la información de la punta del iceberg académico). Obviamente, estos datos procedentes de la zona oculta del iceberg vendrán con un gran ruido documental así como desordenados e imprecisos. Asumir el ruido como algo inherente al big data y trasladar el foco del análisis desde los valores absolutos a nivel micro hacia los clusters y valores relacionales será inevitable.

3. Aplicaciones métricas de Google Scholar

Desde el grupo de investigación EC3 (Evaluación de la Ciencia y de la Evaluación Científica), se pretende analizar Google Scholar en tanto que herramienta válida para la evaluación científica desde una posición neutra. Valorando a través de estudios empíricos sus ventajas e inconvenientes, tanto denunciando sus irregularidades o errores como reconociendo sus prestaciones así como recomendando su buen uso.

Como consecuencia de toda esa actividad, se han elaborado una serie de productos de información basados en los datos proporcionados por Google Scholar, que han permitido aportar tanto evidencias empíricas acerca de su comportamiento y funcionamiento como ofrecer algunas funcionalidades útiles para la comunidad. La descripción breve de los mismos, supone el principal objetivo de este trabajo.

En la figura 1 se ofrece un cuadro resumido de los principales productos a describir. Se distinguen por una parte la familia de productos centrados en exclusiva en una unidad de análisis: los autores (“H Index Scholar”), las editoriales (“Publishers Scholar Metrics”) y las revistas (“Journal Scholar Metrics”). Por otra parte, se propone otra línea de productos basados en la integración de diferentes unidades de análisis en un modelo holístico, en este caso ejemplificado a través de “Biblioteconomía y Documentación Española según Google Scholar Citations”.



Figura 1. Aplicaciones métricas de Google Scholar

3.1. Los autores: H Index Scholar

H Index Scholar (<http://hindexscholar.com>) es ranking de profesores de universidades públicas españolas en las áreas de Ciencias sociales y humanas (Delgado López-Cózar et al, 2014).

Este producto se basa fundamentalmente en la ordenación de los autores en función de su índice H (y complementariamente su índice G) a partir de la producción académica indizada por Google Scholar a fecha de 2012. La principal novedad de este producto es la de ofrecer por primera vez un catálogo de profesores en dichas áreas (se trabajó con una población inicial superior a 40.000 profesores), con el fin de ayudar a reflejar el impacto académico de estos profesionales, generalmente minusvalorados en las bases de datos bibliográficas internacionales. Las características del producto se resumen en la Tabla 1.

CARACTERÍSTICA	DESCRIPCIÓN
Unidad de análisis	Profesores de universidades públicas españolas en Ciencias Sociales y Humanidades
Fuente de datos	Sedes web departamentales más otras fuentes oficiales
Búsqueda de los autores	Búsqueda manual y posterior adscripción a un área de conocimiento
Procesamiento de datos (I)	Búsqueda bibliográfica personalizada

Procesamiento de datos (II)	Cálculo del <i>h-index</i> y <i>g-index</i> .
Visualización	Primer tercil en cada área de conocimiento

Tabla 1. Características del H-Index Scholar (2012)

El proyecto reveló indirectamente grandes inconsistencias en la información ofrecida oficialmente por los sitios web departamentales acerca de los profesores contratados y sus áreas de conocimiento. Así mismo, puso de relieve la necesidad de mejorar la difusión online de la producción científica de los profesores. Por el contrario, el procedimiento arrojó grandes dificultades para desambiguar correctamente los nombres de los autores así como ciertas inconsistencias en algunas áreas (especialmente las relativas a las ciencias jurídicas), que permitieron conocer mejor ciertas debilidades de Google Scholar al respecto. En todo caso, y dado el tamaño de la población analizada, los resultados del proyecto fueron ampliamente satisfactorios.

3.2. Las editoriales: Publishers Scholar Metrics

El ranking de editoriales (<http://www.publishers-scholarmetrics.info>) es un producto derivado de H Index Scholar, a partir de la recolección y posterior identificación de las monografías publicadas por los profesores incluidos en este producto (Tabla 2).

Más allá de la dificultad técnica a la hora de filtrar los datos obtenidos y seleccionar los libros altamente citados, que constituyen la base de este producto (un total de 7.203 monografías), este ranking incluye dos indicadores nuevos, como son el número de libros altamente citados por editorial y el número total de citas recibidas por una editorial, según Google Scholar. Por tanto, el producto ofrece la novedad de obtener datos de impacto a nivel de editorial mediante un procedimiento cuantitativo no basado en encuestas, demostrando la utilidad de Google Scholar para este ámbito.

CARACTERÍSTICA	DESCRIPCIÓN
Muestra inicial	Profesores recogidos en <i>H-index Scholar</i>
Búsqueda de libros (n=1.020.985)	Búsquedas bibliográficas para cada autor
Filtrado de libros (n= 26.628)	Libros no citados Duplicados, erróneos.
Selección de libros (n= 7.203)	Altamente citados
Indicadores	<ul style="list-style-type: none"> Nº de libros altamente citados Nº total de citas de una editorial

Tabla 2. Características del Publishers Scholar Index

3.3. Las revistas: Journal Scholar Metrics

El tercer producto centrado en una unidad de análisis lo constituye el Journal Scholar Metrics, orientado al impacto de las revistas científicas (<http://www.journal-scholar-metrics.infoec3.es>).

En este caso, se trata de un producto en proceso de elaboración actualmente (2015), que pretende recopilar las revistas internacionales en Ciencias sociales y humanas por disciplinas así como ofrecer una serie de indicadores bibliométricos extraídos y elaborados a partir de la información recogida en Google Scholar Metrics. Los detalles del producto se muestran en la Tabla 3. Entre los indicadores elaborados, se debe destacar la incorporación del índice H de las revistas excluyendo las autocitas, dato que no ofrece Google Scholar y que resulta fundamental para la correcta interpretación no sólo del impacto de la revista sino del indicador (robusto pero fuertemente influido por las autocitas).

CARACTERÍSTICA	DESCRIPCIÓN
Identificación de revistas	<i>Google Scholar Metrics, SCImago Journal Rank, Web of Science Master List, Ulrichsweb, etc.</i>
Clasificación de revistas	<ul style="list-style-type: none"> • Ciencias sociales • Humanidades
Indicadores (I): directos	<ul style="list-style-type: none"> • <i>H5-index</i> • <i>H5-median</i> • <i>H5-Citations</i>
Indicadores (II): elaborados	<ul style="list-style-type: none"> • <i>Pure H5-index; Pure H5-Citations</i> • <i>Journal self-citation rate</i>
Visualización	<ul style="list-style-type: none"> • Búsqueda y navegación • Por revista, área y país

Tabla 3. Características del Journal Scholar Metrics

3.4. La biblioteconomía y Documentación Española

Finalmente, se propone un sistema de información holístico a partir de la inclusión de las principales unidades de análisis (autores, documentos, revistas, editoriales e instituciones) para el estudio de una comunidad completa (Biblioteconomía y Documentación), contextualizada en un país (España). En esta ocasión el proceso comienza a partir de la identificación de autores con un perfil abierto en Google Scholar Citations, el servicio de perfiles académicos de Google Scholar. A partir de los autores se obtienen los documentos incluidos en sus perfiles y con la información bibliográfica de los mismos se construye el sistema de información multinivel (<http://www.biblioteconomia-documentacion-española.infoec3.es>).

El potencial de este tipo de análisis radica en la posibilidad de comprender el impacto considerando todos los actores involucrados en el proceso de creación y difusión del conocimiento científico, y no sólo a partir de la calidad de las revistas donde los autores han publicado los resultados de sus trabajos (Figura 2).

**La Biblioteconomía y Documentación española
según Google Scholar Citations**

INICIO / ACERCA DE / METODOLOGÍA / EQUIPO / OTROS PROYECTOS



AUTORES



DOCUMENTOS



REVISTAS



EDITORIALES



INSTITUCIONES

Mostrando revistas 1-25 de 105. Ordenadas por número de documentos, descendientemente.

Nombre de la revista	% de artículos altamente citados	% de citas
El Profesional de la Información	15,9	13,5
Revista Española de Documentación Científica	13,3	10,5
Scientometrics	9,8	14,0
Journal of the Association for Information Science and Technology	7,7	10,4
Anales de Documentación	3,7	3,7
Journal of Informetrics	2,9	6,3
Information Processing & Management	2,9	3,7
Journal of Documentation	2,1	1,2
Information Research	1,6	1,0
Hipertext.net	1,3	1,5
BID: textos universitarios de biblioteconomía y documentación	1,3	1,0
Journal of Information Science	1,1	1,0
Revista Española de Cardiología	1,1	1,0
RUSC: Universities and Knowledge Society Journal	1,1	0,9
Anuario ThinkEPI	1,1	0,9
Revista de Neurología	1,1	0,7
Ciência da Informação	0,8	1,0
The Journal of Academic Librarianship	0,8	0,9
Revista General de Información y Documentación	0,8	0,8
Information Technology and Libraries	0,8	0,8
Medicina Clínica	0,8	0,7
Research Evaluation	0,8	0,6
No Solo Usabilidad	0,8	0,5
Información, Cultura y Sociedad	0,8	0,5
Archivos de Bronconeumología	0,8	0,5

Primera | Anterior | Siguiente | Última

Fecha de actualización: 27/02/2015

Figura 2. Ranking de revistas de Biblioteconomía y Documentación según Google Scholar Citations

La combinación de estos datos con los ofrecidos por la bases de datos bibliográficas elitistas (como WoS) permite conocer mejor el perfil investigador de los autores (más científico o más profesional; orientado a los artículos o a otros canales de comunicación científica).

Del mismo modo, la expansión del análisis de los autores al de los documentos ofrece la posibilidad de conocer qué ideas y temas son los que han generado un mayor impacto e interés dentro de la comunidad. Los canales donde estos documentos se han publicado nos muestran por su parte cuáles son las fuentes (revistas y editoriales) más influyentes en nuestra disciplina. En la Tabla 4 se resumen las principales características de este producto

CARACTERÍSTICA	DESCRIPCIÓN
Unidad de análisis	Autores con perfil público en <i>Google Scholar</i>
Fuente de datos	<i>Google Scholar Citations</i>
Búsqueda e identificación de autores	Palabras clave representativas
Procesamiento de datos	100 documentos más citados de cada autor.
Procesamiento de datos	10% de los documentos altamente citados (percentil 90).

Tabla 4. Características de La Biblioteconomía y Documentación española según Google Scholar Citations

4. Conclusiones

Google Scholar ha experimentado una continua transformación desde su nacimiento en 2004 hasta convertirse actualmente en la mayor base de datos bibliográfica del mundo. El cómputo automático de citas procedentes de cualquier documento académico indizado ha modificado la forma de entender el impacto académico, de forma complementaria a como las Altmetrics lo han logrado, en este caso a través de indicadores no basados en citas (descargas, ReTweets, etc.). La explotación de la información albergada en Google Scholar representa por tanto un enorme desafío a la hora de comprender y valorar el impacto especialmente en áreas en las que las prácticas de publicación no se fundamentan en la publicación de artículos escritos en inglés en revistas. El diseño y elaboración de productos métricos de información como los descritos en este trabajo ponen de manifiesto la posibilidad real existente de trabajar en un entorno complementario al propuesto por las bases de datos bibliográficas clásicas.

Así mismo, el diseño de productos holísticos multinivel abren una ventana hacia la comprensión y contextualización del impacto de una disciplina en su completitud a partir del impacto individual de los diferentes actores que participan en la construcción (autores y documentos) y difusión (revistas y editoriales) de nuevo conocimiento.

Finalmente, la elaboración de estos productos precisa de la comprensión del funcionamiento de Google Scholar, de sus principales ventajas (cobertura, actualización), inconvenientes (control de calidad, opacidad) y posibles amenazas (manipulación). Solamente desde la comprensión profunda de un producto como Google Scholar se pueden elaborar herramientas verdaderamente útiles a la comunidad, principal objetivo de este trabajo.

Referencias bibliográficas

Bar-Ilan, J. (2008). Which h-index?—A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, vol 74, nº 2, p. 257-271.

Delgado López-Cózar, E. (2013). The Google Scholar Family: ¿Is it an alternative for the evaluation of science? En: IV Seminario EC3: Altmetrics y Unidades de Bibliometría. Universidad de Granada, 14-15 marzo. Disponible en: <http://hdl.handle.net/10481/24142>

Delgado-López-Cózar, E.; Orduña-Malea, E.; Jiménez-Contreras, E. y Ruiz-Pérez, R. (2014). H Index Scholar: el índice h de los profesores de las universidades públicas españolas en humanidades y ciencias sociales. *El profesional de la información*, vol. 23, nº 1, p. 87-94.

Harzing, A. W. y Van der Wal, R. (2007). Google Scholar: the democratization of citation analysis. *Ethics in science and environmental politics*, vol. 8, nº 1, p. 61-73.

Harzing, A. W. y Van Der Wal, R. (2009). A Google Scholar h-index for journals: An alternative metric to measure journal impact in economics and business. *Journal of the American Society for Information Science and Technology*, vol. 60, nº 1, p. 41-46.

Jacsó, P. (2008). Google scholar revisited. *Online information review*, vol. 32, nº 1, p. 102-114.

Jacsó, P. (2009). Calculating the h-index and other bibliometric and scientometric indicators from Google Scholar with the Publish or Perish software. *Online Information Review*, vol. 33, nº 6, p. 1189-1200.

Jacsó, P. (2010). Metadata mega mess in Google Scholar. *Online Information Review*, vol. 34, nº 1, p. 175-191.

Jacsó, P. (2011). The pros and cons of Microsoft Academic Search from a bibliometric perspective. *Online Information Review*, vol. 35, nº 6, p. 983-997.

Kousha, K. y Thelwall, M. (2007). Google Scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology*, vol. 58, n. 7, p. 1055-1065.

Orduna-Malea, E. y López-Cózar, E. D. (2014). Google Scholar Metrics evolution: an analysis according to languages. *Scientometrics*, vol. 98, n. 3, p. 2353-2367.

Orduna-Malea, E.; Ayllón, J. M.; Martín-Martín, A. y Delgado López-Cózar, E. (2015). Methods for estimating the size of Google Scholar. *Scientometrics*, vol. 104, n. 3, p. 931-949.

Orduna-Malea, E.; Martín-Martín, A.; Ayllon, Juan M. y Delgado Lopez-Cozar, E. (2014). The silent fading of an academic search engine: the case of Microsoft Academic Search. *Online Information Review*, vol. 38, nº 7, p. 936-953.

Ortega, Jose L. (2014). *Academic search engines: A quantitative outlook*. Elsevier.

Torres-Salinas, D.; Ruiz-Pérez, R. y Delgado-López-Cózar, E. (2009). Google Scholar como herramienta para la evaluación científica. *El profesional de la información*, vol. 18, nº 5, p. 501-510.

Van Noorden, R. (2014). Online collaboration: Scientists and the social network. Nature, vol. 512, nº 7513, p. 126-129.

Esta obra se distribuye bajo licencia Creative Commons (CC) 3.0, disponible en:

<http://creativecommons.org/licenses/by-nc/3.0/es/deed.es> AR

